Exploring Methods for Improving the Integration of LOINC in the UMLS

Holli Huff[a], Olivier Bodenreider, MD PhD[b]

[a] *Department of Medical Informatics, University of Utah, Salt Lake City, UT*
[b] *US National Library of Medicine, National Institutes of Health, Dept. of Health & Human Services, Bethesda, MD*

Abstract

The UMLS Metathesaurus has included the Logical Observation Identifier Names and Codes (LOINC) medical coding system into it's content. However, although it is present, it is not well integrated. There are very few relationships between LOINC concepts and other Metathesaurus concepts. This paper describes a method for detecting these relationships. Two methods were used, a generative and analytical, for discovering patterns in the natural language target terms. The analytical approach had much better results and determined the structure of 4543 unique concepts. Once this structure is known, it can be used to discover similarities between LOINC and natural language terms. These similarities infer that relationships exist between the terms. Although only preliminary work has been done in this regard, the authors predict that between 10,000 and 100,000 relationships can be detected in this manner.

Introduction

Laboratory Observation Identifier Names and Codes (LOINC) is a medical coding system used for the identification of laboratory and clinical observations. The laboratory observations are the most widely adopted and will be the focus of this paper. LOINC is widely supported and has been funded in part by NLM, HCFA, DOD, AHRQ, and the John A. Hartford Foundation.[1] It is publicly available at http://www.loinc.org. It is very important that LOINC be well incorporated into the Metathesaurus. Unfortunately this is not the case.

Concepts in the UMLS Metathesaurus generally have one of two structure types. They are either natural language or fielded data. LOINC terms are in the form of fielded data. These two structure types are used to encode similar or identical information in very different ways. This causes the integration of LOINC to be very difficult. A method for integrating the two structure types is needed.

Figure 1 contains comparisons of natural language strings to LOINC strings. Notice that these two pairs have things in common. In the first, "serum" – "SER" and "creatinine" are common. In the second, "urine" - "UR", "cannabinoids", and "screen" are common. Although these terms have common constituents, they are in very different forms.

Figure 1 Contrast in Forms of Natural Language and LOINC Concepts

Serum Creatinine Tests	natural language
CREATININE:SCNC:PT:SER/PLA:QN	LOINC
Urine Cannabinoids Screen	natural language
CANNABINOIDS:ACNC:PT:UR:ORD:SCREEN	LOINC

Fully specified LOINC names are composed of 6 axes separated by colons. They are component/analyte, property, timing, system, scale, and method (optional). Often, LOINC names consist mainly of abbreviations.[2]

The objective of this project was the creation of a method to facilitate the detection of missing relationships (including synonymy) between natural language and LOINC terms in the UMLS. Although this end has not yet been met, the groundwork has been done. Some preliminary results are included herein.

Methods

There were three phases to this project. They were 1) identify the natural language terms that could possibly have any relationship to LOINC observations (Natural Language Target Terms), 2) identify patterns in the target terms (Generative and Analytical Approaches), and 3) match similar LOINC and natural language terms based on their constituents. Each phase will be discussed in detail.

Natural Language Target Terms

Natural language terms were retrieved from the UMLS Metathesaurus based on two different techniques. One was based on semantic types, the other on hierarchies. 3 semantic types were chosen which seemed likely to be applied to concepts that would have relationships with LOINC observations. They were T201 Clinical Attribute, T059 Laboratory Procedure, and T034 Laboratory or Test Result. The majority of LOINC observations are typed as Clinical Attributes. The other are both laboratory related and seemed like they could have possible overlap with LOINC meanings. All of the concepts with these semantic types were retrieved from the Metathesaurus. They consisted of 10,788 concepts and 20,903 strings. The second technique was based on the hierarchical structure of UMLS. All of the descendents of C0427351 Laboratory

Test Observations and C0600201 Laboratory Techniques and Procedures were retrieved for this purpose. This retrieval yielded 54,792 concepts and 131,744 strings. The first list is much smaller because it is a cleaner way to retrieve concepts from the Metathesaurus. The hierarchical relationships are not precise and therefore many things were retrieved which had no relationships to LOINC observations. However, this was not of concern because this list needed to be all-inclusive and specificity was not a concern.

After compiling the lists of target terms two approaches were taken for identifying patterns in the natural language strings; a generative approach and an analytical approach. Each is described in detail in the following paragraphs.

Generative Approach

The generative approach began with the construction of a grammar by inspection. Symbols such as <ANA> for analyte and <SYS> for system were used to represent the parts of speech. For example, the natural language term "serum sodium" would be represented as "<SYS> <ANA>". Next, a lexicon was constructed to include any possible values found in the LOINC axes. Since LOINC names contain many abbreviations, the expanded LOINC names from the Metathesaurus were included in the lexicon. There are also strings, which are not values in the LOINC axes. For example, "test" or "decreased". These strings were categorized and manually added to the lexicon. Next, rules of the grammar were combined with each string in the lexicon to form natural language string. Lastly, string matching was used to test the completeness of the grammar and lexicon. This process was iterated 3 times in order to make the grammar and lexicon more robust.

Analytical Approach

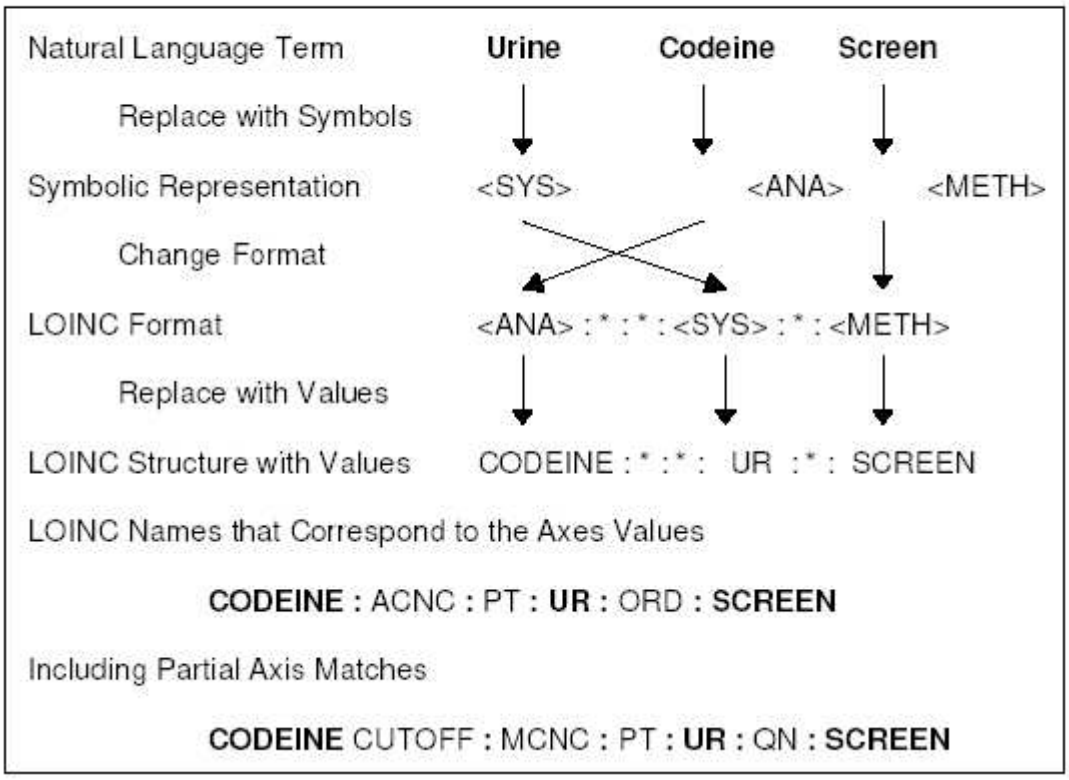
The analytical approach began with the list of natural language target terms. The same symbols for the parts of speech were used as in the generative approach. A Perl script was utilized to read the natural language terms, refer to the lexicon and replace each term with it's appropriate symbol. After all of the natural language terms were processed, they were inspected for grammatical patterns. Each natural language string that was completely replaced by symbols was a completed pattern. The terms that were only partially

matched were inspected and frequently occurring terms were added to the lexicon. This approach was also iterative and was completed twice.

Matching Natural Language Strings to Existing LOINC Names

This part of the project is in very preliminary stages, but it has great potential to produce meaningful results. The detection of relationships is best described by example. Please refer to figure 2. The process begins with a natural language term. Each concept is replaced with its corresponding symbol from the grammar. The symbols are then mapped to their positions in the specific LOINC structure. The LOINC values are read from the lexicon and replace the symbols. Finally, this representation is used to match LOINC names that have common constituents. Finding these "matching concepts" will be very beneficial in detecting where relationships exist and what type they are.

Figure2 Process for Detecting Relationships



Results

The generative approach led to the discovery of 19 grammar rules and the third iteration produced 2.4 billion strings consuming 111 gigabytes of disk space. The string generation and matching programs took approximately 50 hours to run. 5252 strings representing 1637 unique concepts were matched. The second execution led to the discovery of many more patterns than the generative approach (458). It also ran faster and consumed much less disk space. There were 11,134 completely replaced natural language target terms, representing 4543 unique concepts. Thus, the analytical approach had much better results by identifying the structure of almost 3000 more concepts than did the generative approach.

Discussion

The results of these two approaches are hard to interpret because they are only a stepping-stone to the actual objective, detecting relationships. It cannot be determined if the analytical approach was effective until more research has been done on the detection of relationships.

Figure 3 Matching

Urine Opiates Screen

<ANA>	*	*	<SYS>	*	<METH>
OPIATES	ACNC	PT	**UR**	ORD	**SCREEN**
OPIATES	ACNC	PT	**UR**	ORD	**SCREEN**>2000 NG/ML
OPIATES	ACNC	PT	**UR**	ORD	SAMHSA **SCREEN**
OPIATES CUTOFF	MCNC	PT	**UR**	QN	**SCREEN**
OPIATES TESTED FOR	PRID	PT	**UR**	NAR	**SCREEN**
OPIATES TESTED FOR	PRID	PT	**UR**	NOM	**SCREEN**

Figure 3 shows another example of the types of matches that can be found between natural language terms and LOINC names. The first row shows the symbols representing the pattern of the natural language term.

The second row is a LOINC name with exact matches for the first, fourth, and sixth axes. The third and fourth rows show LOINC names that are only partial matches for the sixth axis. These are more specific types of screens performed and would then be children with the second row as its parent. The last three rows are partial matches on the first axis and take a slightly different twist on "opiates". Relationships also exist here, but their type needs to be explored. Since the sixth axis, method, is optional not specifying it leads to broader LOINC names. If no restraints are put on method in this scenario, 11 more LOINC names become legitimate matches to the original natural language term and they also have relationships to be determined. Obviously there is much work to be done here. Finding the patterns that these relationships follow will be very interesting future work.

Acknowledgements

This project was completed as part of a two-month summer rotation at the National Library of Medicine. Special thanks are offered to Alexa McCray for providing the opportunity and to Tom Rindflesch and May Cheh for supervising it.

References

[1] http://www.loinc.org/background
[2] http://www.loinc.org/download/loinc/guide/LOINCManual_200305.pdf